Speech Emotion Recognition Using Auditory Spectrogram and Cepstral Features

Shujie Zhao[1], Yan Yang[1*], Israel Cohen[2], and Lijun Zhang[1]
[1]: CIAIC, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
[2]: Technion–Israel Institute of Technology, Haifa 3200003, Israel

***Abstract*—A systematic comparison on the impact of environmental noises on key acoustic features is critical in order to transfer speech emotion recognition (SER) systems into real world applications. In this study, we investigate the noise-tolerance of different acoustic features in distinguishing various emotions by comparing the SER classification performance on clean speech signals and noisy speech signals. We extract the spectrum and cepstral parameters based on human auditory characteristics and develop machine learning algorithms to classify four types of emotions using these features. Experimental results across the clean and noisy data show that compared to cepstral features, the auditory spectrogram-based features can achieve higher recognition accuracy for low signal-to-noise ratios (SNRs), but lower accuracy for high SNRs. Gammatone filter cepstral coefficients (GFCCs) outperformed all the extracted features on the Berlin database of emotional speech (EmoDB), under all four kinds of tested noise conditions. These results show compensation relationships between auditory spectrogram-based features and cepstral features for SER with better noise robustness in real-world applications.**

***Index Terms*—Emotion recognition, speech signals, machine learning, pattern recognition, feature extraction, noise**

I. Introduction

Emotional aspect of speech is an important factor in human communication. Through an accurate perception of people's emotions from their speech, human beings can achieve effective interpersonal communication. Emotion recognition via speech focuses on automatically identifying the affective state of a person from speech and has many applications, e.g., detecting potential problematic points causing anger and frustration in call centers [1], recognizing stress response to a stimuli (questions) in lie detection [2], and detecting uncertainty and confidence of students in spoken dialogue computer tutors [3]. In natural environments, the desired signals for speech emotion recognition (SER) usually coexist with background noises. Thus, algorithms developed on clean speech signals have insufficient recognition accuracy in real-world noisy condition. Signal acquisition in typical indoor environments is characterized by echo, reverberation, interference and additive noise, which all lead to degradation of the quality and reliability of speech related recognition tasks [4]–[8]. It is therefore necessary to address the noise robustness problem for SER in real-world applications.

This project was supported in part by the National Natural Science Foundation of China (NSFC) under grant no. 6177012290 (F011305). The work of S. Zhao was supported in part by the China Scholarship Council.

*: Corresponding author

Human ears have a good anti-noise recognition ability. Therefore, in the speech recognition domain, many studies have been devoted to the auditory characteristics of human ears, and many signal processing approximation methods were proposed to simulate the frequency-domain analysis methods of human ears, so as to establish the voice feature parameter model more in line with the auditory characteristics of human ears [9], [10]. Some emotional speaker recognition results show that auditory features can improve the speech recognition results and enhance the noise robustness of the system [11]–[13]. Moreover, some novel features based on human peripheral hearing system were extracted to increase the robustness of SER in noise and reverberation scenarios, such as features based on supervised nonnegative matrix factorization (NMF) [14]–[16], damped oscillator cepstral coefficients (DOCCs) [17], Teager energy cepstrum coefficient (TECC) [18], cochlear filterbanks [19], [20], and pooling scheme-based modulation [21].

Extracting the effective emotion information from raw audio data is still an open challenge. A systematic understanding of the noise robust feature representation for emotional speech is fundamentally indispensable. In this study, we investigate spectrum and cepstral parameters based on human auditory characteristics. We present a comprehensive comparison of the four kinds of spectrum and cepstral parameters on their SER performances across clean data and artificial additive noise data. Simulation results show that auditory spectrogram-based features yield a more robust performance than cepstral features under lower signal-to-noise ratio (SNR) conditions, while cepstral features are advantageous for higher SNRs.

II. Data Collection

EmoDB is a German open database, including 10 actors (5 female and 5 male) and 7 types of emotions (neutral, anger, fear, joy, sadness, disgust, and boredom). The data were gathered in an anechoic chamber with a sampling rate of 16kHz. In the expert testing phase, the emotional utterances that were rated higher than 80% and non-emotional utterances (i.e. naturalness) higher than 60% were retained. Finally, the database includes 535 utterances [22]. Here, we only include the speech with happy (joy), angry, neutral and sad emotions for further analysis.

By overlaying the clean speech signals from EmoDB with Gaussian white noise, pink noise, factory noise, and vehicle noise from Noise-92 database, we generate four extra

artificial additive data sets which simulated acted emotions in the presence of background noise. This data is used for evaluating the recognition performance of different speech features in different noise environments under various levels of SNRs (i.e., from -10 to 40dB with increments of 5dB), in comparison with recognition performance without noise.

III. Feature Extraction

A. Mel filter

Mel frequency cepstral coefficients (MFCCs) proposed by Davis and Mermelstein [23], were based on the sensitivity of human perception to frequency and were popularly used in automatic speech and speaker recognition [24]. Based on the Mel perceptive frequency scale [25], the cosine transform of the short-time power spectrum in the real logarithmic domain is firstly calculated, and then MFCCs are obtained by applying the discrete cosine transform (DCT) to the Mel-filter banks.

B. Cochlear filter

On the basis of simulating the basement membrane response of the human ear, the cochlear filter realizes the whole process of sound transmission from the outer ear to the basement membrane through wavelet transform, called auditory transformation, which is defined as [26]

$$T(a,b) = \int_{-\infty}^{+\infty} x(t)\psi_{a,b}(t)dt = x(t) * \psi_{a,b}(t) \tag{1}$$

where $x(t)$ is the speech signal in time domain, $\psi_{a,b}(t)$ is the cochlear filter function, $*$ denotes a convolution operation.

$$\begin{aligned}\psi_{a,b}(t) &= \frac{1}{\sqrt{a}}\psi(\frac{t-b}{a}) \\ &= \frac{1}{\sqrt{a}}(\frac{t-b}{a})^{\alpha}\text{exp}(-2\pi f_L\beta(\frac{t-b}{a}) \\ &\quad \cos(2\pi f_L\beta((t-b)/a+\theta))u(t-b)\end{aligned} \tag{2}$$

where α and β are real numbers greater than zero. They together control the shape and width of $\psi_{a,b}(t)$ in frequency domain. In this study, their values were set as 0.3 and 0.2, since the frequency response curve of cochlear filter bank is closer to the auditory frequency response curve of human ear. The function $u(t)$ is the unit step function, θ is the initial phase, and b is a time varying real value. The parameter a is a scale variable determined by the central frequency f_c and the lowest central frequency f_L of the filter bank, $1/\sqrt{a}$ is an energy normalization factor, which ensures a consistent energy under various a and b values. Sensory hair cell have the function to convert the auditory speech signal into the nerve pulse signal. The hair cell window uses different window lengths to analyze different signals, due to the fact that the nerve pulse signal generated by different frequency signals are not the same. Then the energy information of the acquired signal is transformed into perceived loudness through a nonlinear loudness transformation function. Finally, the cochlear filter cepstral coefficients (CFCCs) can be obtained through DCT.

C. Gammatone filter

In 1992, Roy Patterson and his colleagues designed a Gammatone filter based on the frequency response in the basement membrane of human ears. By simulating the traveling waves in the basement membrane of the cochlea, the time-domain speech signal was decomposed into a series of frequency band information. The impulse response of the Gammatone filter bank is defined as [13]:

$$g_i(t) = A\,t^{n-1}\text{exp}(-2\pi b_i t)\text{cos}(2\pi f_i + \phi_i)u(t) \tag{3}$$

where, $t \geq 0$, and $1 \leq i \leq N$, A is the filter gain, n is the order, f_i is the center frequency, $u(t)$ denotes the step function, N is the number of filters, and ϕ_i is the initial phase. The parameter b_i represents the attenuation factor determining the attenuation rate of the impulse response and can be depicted as

$$b_i = 1.019\,b_{\text{ERB}}(f_i) \tag{4}$$

where $b_{\text{ERB}}(f_i)$ is the equivalent rectangular bandwidth (ERB) of each filter, which is related to the center frequency of the filter and the critical frequency band of human auditory system. The value of $b_{\text{ERB}}(f_i)$ in auditory psychology model is given as

$$b_{\text{ERB}}(f_i) = 24.7(4.37 \times \frac{f_i}{1000} + 1) \tag{5}$$

The central frequency is proportional to the bandwidth on a logarithmic scale, that is, it has a nonlinear frequency characteristic and conforms to the auditory characteristics of human ears.

Here, the number of mel filters is 24 and cochlear filters is 18, and for gammatone filters, the order of the filters was set as 4 and the number of filters was empirically set to 64. The sampling frequency is 16kHz, frame size is 25ms, the frequency response curves of these filter banks obtained by triangular shaped filters are shown in Fig. 1.

D. Log-Spectrum

We adopt the signal spectrogram to define the log-spectrum feature parameters [27]:

$$S(i) = \frac{1}{M}\sum_{m=1}^{M}\log|X(m,i)| \tag{6}$$

where, i represents the frequency band index, M is the number of frames contained in a utterance, and $X(m,i)$ denotes the discrete Fourier transform of the signal in the mth frame. In the experiment, we refer to the previous work and analyze the information within frequency interval of 0–1200 Hz, which corresponds to the low frequency component, namely the first 30 mean of log-spectrum (MLS) coefficients [28]. In addition to MLS coefficients, we also extract some auditory spectrogram coefficients, as an extension of spectrum features, such as robust MFCC (MFCC-R) [27], robust CFCC (CFCC-R), and robust GFCC (GFCC-R). It is worth noting that CFCC-R and GFCC-R features have never been used before for SER in noisy conditions.

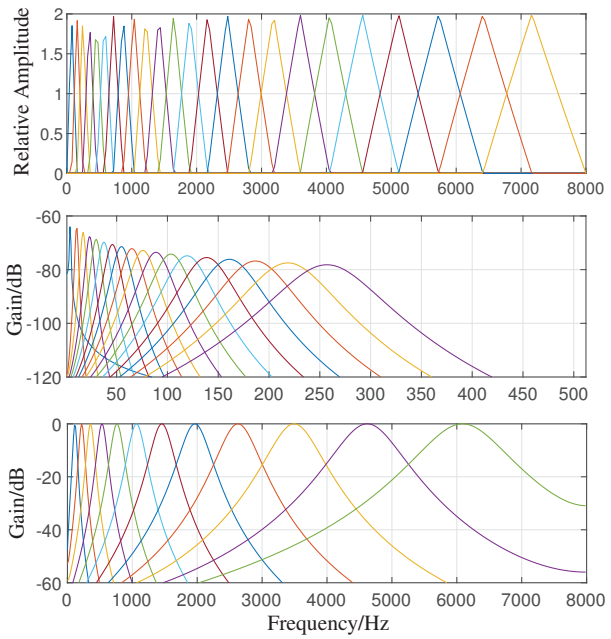

Fig. 1. Frequency response curve of mel-filter-bank (top), cochlear-filter-bank (middle) and gammatone-filter-bank (bottom).

TABLE I
FEATURE SET.

Feature groups	MFCC	MFCC-R	MLS	CFCC	CFCC-R	GFCC	GFCC-R
No. dimension	37	37	30	54	18	93	64

In order to satisfy the structural features of human ears and highlight the dynamic changes of speech signals, the first 12 MFCCs, the first 18 CFCCs, the first 31 GFCCs and their zero-order difference coefficients, first-order difference coefficients and second-order difference coefficients were extracted based on speech frames. At last, the mean of each cepstral and spectrum coefficients was calculated for each utterance. All the features used in our experiments are shown in TABLE I.

IV. NOISE ROBUST EXPERIMENTS

In this section, the robustness of the features mentioned before for SER were evaluated based on the back propagation (BP) neural network classifier. The number of nodes in the input and output layers were determined by the input and output sequences of the SER task. The connection weights between neurons in the input layer, hidden layer and output layer were randomly initialized. The bias of neurons in the hidden layer and the output layer were also randomly initialized. The learning rate was set to 0.1. The log sigmoid function was used as the nonlinear activation function and the additional momentum method in MATLAB Toolbox was adopted to update the weights. The data was first normalized to the scale of $[0, 1]$, and then partitioned using a triangular shaped window into frames of 400 samples, with a frame shift of 160 samples. With an eye to extracting real voice utterances and reducing the computational burden of subsequent processing, endpoint detection was also performed through each utterance. After preprocessing, the feature parameters were extracted as it was described in Section III. During training, 75% data from the extracted features used as the training data and the remaining 25% as the test data. The split of training and testing data was randomized. The trained network that achieved the best test results on clean data was also used to evaluate additive noise data.

TABLE II
EMOTION RECOGNITION RESULTS OF DIFFERENT FEATURES ON EMODB. ACCURACY IN [%].

	Anger	Happiness	Neutral	Sadness	Average
MFCC	92.59	71.43	85.71	100.00	**87.43**
MFCC-R	96.30	73.68	84.21	95.00	87.30
MLS	90.32	70.59	94.44	89.47	86.21
CFCC	100.00	73.68	94.74	88.24	**89.16**
CFCC-R	90.00	71.43	87.50	100.00	87.23
GFCC	92.86	83.33	100.00	93.33	**92.38**
GFCC-R	94.29	77.78	92.86	100.00	91.23

A. Experiments on EmoDB

As we can see from the results in TABLE II, for clean speech, the average recognition accuracy rate based on GFCC is higher than that those on MLS by 6.17%, MFCC by 4.95% and CFCC by 3.22%. In addition, the average recognition accuracy rate with cepstral coefficients and their difference coefficients is respectively higher than that of spectrum coefficients. For example, taking relative error as an index, the accuracy rate of MFCC surpass MFCC-R by 0.14%, CFCC is over CFCC-R by 2.21%, and GFCC is higher than GFCC-R by 1.26%. Cepstral coefficients have some advantages. First, in terms of algorithm, the feature extraction process of cepstral coefficients includes discrete cosine transform (DCT), which has the advantage of sparse signal spectral components and concentrated energy. DCT can also achieve a good speech enhancement result with low computational complexity, while the speech enhancement process can improve the SER performance of feature parameters in a sense. Secondly, since speech signal is a short-term stationary signal, most researchers extract emotional features by frame processing of the speech signal. However, the features extracted based on a certain frame are local features, which cannot accurately reflect the dynamic characteristics of emotional speech. Therefore, it is often impossible to build a robust emotional recognition system by simply adopting local features. After framing, by extracting the differential parameters of local features at the statement level and fusing the two statement-level features together, the classification performance can be effectively improved. The experimental results based on EmoDB show that the fused static and difference features improve the recognition rate of locally static features by 1.65% for MFCC, 1.25% for CFCC and 2.57% for GFCC.

TABLE III
EMOTION RECOGNITION RESULTS UNDER GAUSSION WHITE NOISE CONDITION. ACCURACY IN [%].

	MFCC	MFCC-R	MLS	CFCC	CFCC-R	GFCC	GFCC-R
-10dB	25.40	25.06	24.04	25.00	24.57	22.58	**46.12**
-5dB	26.21	26.56	**54.41**	25.81	29.05	25.37	41.53
0dB	28.23	29.94	**76.15**	31.59	27.07	32.04	27.82
5dB	34.52	30.26	**70.58**	46.70	26.57	48.08	29.15
10dB	46.21	43.46	**72.75**	62.30	54.36	61.68	45.06
15dB	56.25	60.78	78.98	76.82	76.48	**79.16**	61.35
20dB	61.98	72.92	79.76	80.40	85.95	**88.44**	79.38
25dB	64.02	77.99	80.27	83.33	88.65	**92.09**	86.84
30dB	68.45	81.50	80.27	84.29	90.20	**94.03**	89.05
35dB	79.82	83.75	80.47	85.36	90.60	**94.11**	89.48
40dB	83.45	84.34	80.82	86.11	90.72	**94.71**	90.15

TABLE IV
EMOTION RECOGNITION RESULTS UNDER PINK NOISE CONDITION. ACCURACY IN [%].

	MFCC	MFCC-R	MLS	CFCC	CFCC-R	GFCC	GFCC-R
-10dB	25.00	23.20	**36.69**	29.11	25.40	28.00	24.19
-5dB	26.97	24.78	**44.07**	34.68	25.86	27.98	22.21
0dB	41.93	28.34	**62.98**	35.09	31.50	31.12	28.33
5dB	49.02	39.97	**70.35**	56.30	51.36	40.60	32.10
10dB	51.87	53.58	**72.50**	63.71	67.16	63.11	46.78
15dB	55.94	63.28	76.81	80.64	**84.31**	82.11	60.80
20dB	60.18	69.14	80.46	81.96	87.92	**91.87**	81.33
25dB	67.24	76.00	80.26	84.73	89.65	**92.16**	89.41
30dB	77.84	79.12	80.47	85.63	89.61	**94.58**	91.45
35dB	84.98	82.35	80.47	86.34	89.52	**94.77**	90.42
40dB	87.00	82.18	81.17	87.09	91.10	**95.02**	90.89

TABLE V
EMOTION RECOGNITION RESULTS UNDER FACTORY NOISE CONDITION. ACCURACY IN [%].

	MFCC	MFCC-R	MLS	CFCC	CFCC-R	GFCC	GFCC-R
-10dB	25.00	22.39	34.73	25.00	25.00	26.59	**50.94**
-5dB	25.39	25.47	34.35	25.00	27.82	33.30	**67.65**
0dB	30.78	43.26	42.54	25.00	43.96	44.96	**79.81**
5dB	48.88	59.93	68.06	30.00	34.15	62.35	**84.30**
10dB	67.45	70.07	69.29	42.50	40.98	77.08	**88.31**
15dB	78.71	75.80	75.93	62.04	59.04	87.99	**90.02**
20dB	83.78	79.12	80.03	80.33	77.58	**91.10**	90.96
25dB	86.44	82.00	79.59	87.95	85.74	**94.42**	90.61
30dB	88.64	82.18	79.15	87.79	88.43	**94.82**	90.26
35dB	90.44	82.22	79.39	90.24	89.77	**94.86**	90.85
40dB	91.19	82.06	80.37	90.15	90.90	**94.75**	90.49

TABLE VI
EMOTION RECOGNITION RESULTS UNDER VEHICLE NOISE CONDITION. ACCURACY IN [%].

	MFCC	MFCC-R	MLS	CFCC	CFCC-R	GFCC	GFCC-R
-10dB	28.48	**53.01**	25.00	25.00	25.00	26.59	50.94
-5dB	34.79	**71.56**	28.04	25.00	27.82	33.30	67.65
0dB	50.41	76.66	45.88	25.00	43.96	44.96	**79.81**
5dB	69.68	80.09	64.54	30.00	34.15	62.35	**84.30**
10dB	77.82	81.99	67.11	55.97	40.98	77.08	**88.31**
15dB	83.09	82.18	75.70	62.04	59.04	87.99	**90.02**
20dB	87.37	82.57	78.84	80.33	77.58	**91.10**	90.96
25dB	90.79	82.37	80.14	87.95	85.74	**94.42**	90.61
30dB	91.42	82.76	80.39	87.79	88.43	**94.82**	90.26
35dB	91.11	83.11	79.61	90.24	89.77	**94.86**	90.85
40dB	92.44	83.11	79.65	90.15	90.90	**94.75**	90.49

B. Experiments on artificial additive data

The robust classification performance of speech features is tested on the artificial additive noise dataset. In the experiments, the classifier was trained with clean speech signals and then tested with noisy speech signals from other speakers, to avoid speaker dependency. The experimental results on noisy data are shown in Tables III, IV, V, and VI.

From Tables III–VI, we can see that the average recognition rates of almost all features drop off along the decrease of SNRs. Under lower SNRs, auditory spectrogram-based coefficients are more significant to SER than cepstral coefficients. For example, under Gaussion noise condition, MLS yields the highest accuracy among all other features at 10 dB, 5 dB, 0 dB, and -5 dB. Under factory noise condition, GFCC-R achieves a higher accuracy than other features below 20 dB. The advantageous property of auditory spectrogram-based coefficients can also be noticed under pink noise and vehicle noise conditions. under higher SNRs conditions, cepstral features improve the performance more than auditory spectrogram-based features. Taking GFCC for example, GFCC achieves the best accuracy rate among all the tested features For SNR above 20 dB and for for all the four kinds of noise. This phenomenon is in accordance with the results on EmoDB, where GFCC has the highest recognition rate over all features and the difference value of recognition accuracy based on different features is not greater than 0.5. Though cepstral features, such as MFCC, are sub-band energy-based features, which have good representations of speech spectral information, they are sensitive to noise and thus are less useful for identifying emotions from noisy speech.

V. CONCLUSIONS

In this paper, we have conducted a comprehensive comparative study of the auditory spectrogram-based features and cepstral features on SER with two types of speech data sources (clean speech, and noisy speech). Experimental results show that auditory spectrogram-based features yield a more

robust performance than cepstral features under lower SNRs. Cepstral features, on the other hand, are able to improve the performance in terms of classification accuracy under higher SNRs, more than auditory spectrogram based features. In future work, we plan to further investigate the compensation relationship between auditory spectrogram-based features and cepstral features when coping with emotion recognition tasks, and design a fusion scheme for SER with better noise robustness.

REFERENCES

[1] D. Morrison, RL. Wang, and LC. De Silva, "Ensemble methods for spoken emotion recognition in call-centres,", Speech Communication, vol. 49, no. 2, pp. 98-112, 2007.

[2] F. Horvath, "Detecting deception: the promise and the reality of voice stress analysis," Journal of Forensic Sciences, vol. 27, no. 2, pp. 340-351, 1982.

[3] K. Forbes-Riley, and D. Litman, "Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor," Speech Communication, vol. 53, no. 9-10, pp. 1115-1136, 2011.

[4] H. W. Loellmann, H. Barfuss, A. Deleforge, S. Meier, and W. Kellermann, "Challenges in acoustic signal enhancement for human-robot communication," Speech Communication, pp. 1-4, 2014.

[5] S. Zhao, Y. Yang, and J. Chen, "Effect of reverberation in speech-based emotion recognition," 2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE), Eilat, Israel, 2018, pp. 1-5.

[6] J. Pohjalainen, F. Ringeval, Z. Zhang, and B. Schuller, "Spectral and cepstral audio noise reduction techniques in speech emotion recognition," Proceedings of the 24th ACM Multimedia Conference, 2016, pp. 670-674.

[7] M. Pandharipande, R. Chakraborty, A. Panda, and S. K. Kopparapu, "An unsupervised frame selection technique for robust emotion recognition in noisy speech," 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 2018, pp. 2055-2059.

[8] A. K. Alimuradov, A. Y. Tychkov and P. P. Churakov, "A method for noise-robust speech signal processing to assess human psycho-emotional state," 2019 3rd School on Dynamics of Complex Networks and their Application in Intellectual Robotics (DCNAIR), Innopolis, Russia, 2019, pp. 6-8.

[9] MCA. Korba, H. Bourouba, and R. Djemili, "Feature extraction algorithm using new cepstral techniques for robust speech recognition," Malaysian Journal of Computer Science, vol. 33, no. 2, pp. 90-101, 2020.

[10] U. Kumaran, SR. Radha, SM. Nagarajan, and A. Prathik, "Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN," International Journal of Speech Technology, Jan 2021.

[11] M. Russo, M. Stella, M. Sikora, and V. Pekic, "Robust cochlear-model-based speech recognition," Computers, vol. 8, no. 1, Jan 2019.

[12] A. Mansour, and Z. Lachiri, "A comparative study in emotional speaker recognition in noisy environment," 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications, Hammamet, 2017, pp. 980-986.

[13] X. Zhao, Y. Shao, and D. Wang, "CASA-Based robust speaker identification," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 5, pp. 1608-1616, July 2012.

[14] F. Weninger, B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognition of nonprototypical emotions in reverberated and noisy speech by nonnegative matrix factorization," Eurasip Journal on Advances in Signal Processing, 2011.

[15] MX. Hou, JX. Li, and GM. Lu, "A supervised non-negative matrix factorization model for speech emotion recognition,", Speech Communication, vol. 124, pp. 13-20, Nov 2020.

[16] SR. Bandela, and TK. Kumar, "Unsupervised feature selection and NMF de-noising for robust speech emotion recognition," Applied Acoustics, vol. 172, Jan 2021.

[17] V. Mitra, A. Tsiartas, and E. Shriberg, "Noise and reverberation effects on depression detection from speech," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, 2016, pp. 5795-5799.

[18] R. Sun, and E. Moore II, "Investigating the robustness of teager energy cepstrum coefficients for emotion recognition in noisy conditions," Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference (FLAIRS), 2012.

[19] P. K. Aher, S. D. Daphal, A. N. Cheeran, "Analysis of feature extraction techniques for improved emotion recognition in presence of additive noise," 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India, 2016, pp. 350-354.

[20] S. Hamsa, I. Shahin, Y. Iraqi, and N. Werghi, "Emotion recognition from speech using wavelet packet transform cochlear filter bank and random forest classifier," IEEE Access, vol. 8, pp. 96994-97006, 2020.

[21] A. R. Avila, J. Monteiro, D. O'Shaughneussy, and T. H. Falk, "Speech emotion recognition on mobile devices based on modulation spectral feature pooling and deep neural networks," 2017 in IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Bilbao, 2017, pp. 360-365.

[22] F. Burkhardt, A. Paeschke, M. Rolfes, WF. Sendlmeier, and B. Weiss, "A Database of German emotional speech," Proceedings of 9th European Conference on Speech Communication and Technology (Interspeech), Lisbon, Portugal, Sep 2005, DBLP, 2005, pp. 1517-1520.

[23] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 28, no. 4, pp. 357-366, Aug 1980.

[24] F. Albu, D. Hagiescu, L. Vladutu, and M.A. Puica, "Neural network approaches for children's emotion recognition in intelligent learning applications," EDULEARN 15 7th Annu Int Conf Educ New Learn Technol Barcelona, Spain, 6th-8th, 2015.

[25] Stevens, and S. S., "A scale for the measurement of the psychological magnitude pitch," Journal of the Acoustical Society of America, vol. 8, no. 3, pp. 185-190, 1937.

[26] Q. Li and Y. Huang, "An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 6, pp. 1791-1801, Aug 2011.

[27] EM. Albornoz, DH. Milone, and HL. Rufine, "Feature extraction based on bio-inspired model for robust emotion recognition," Soft Computing, vol. 21, no. 17, pp. 5145-5158, Sep 2017.

[28] E. M. Albornoz, D. H. Milone, and H. L. Rufine, "Spoken emotion recognition using hierarchical classifiers," Computer Speech and Language, vol. 25, no. 3, pp. 556-570, Jul 2011.